Coordinating Government Funding of File System and I/O Research through the High End Computing University Research Activity

Gary Grider, James Nunez, John Bent
Los Alamos National Lab
DOE/NNSA
{ggrider, jnunez, johnbent}@lanl.gov

Rob Ross
Argonne National Lab
DOE/Office of Science
rross@mcs.anl.gov

Lee Ward
Sandia National Lab
DOE/NNSA
lee@sandia.gov

Steve Poole
Oak Ridge National Lab
DOE/Office of Science
spoole@ornl.gov

Evan Felix
Pacific Northwest National Lab
DOE/Office of Science
Evan.Felix@pnl.gov

Ellen Salmon
NASA
Ellen.M.Salmon@nasa.gov

Marti Bancroft
Department of Defense/NRO
marti@dragonsden.com

ABSTRACT
In 2003, the High End Computing Revitalization Task Force designated file systems and I/O as an area in need of national focus. The purpose of the High End Computing Interagency Working Group (HECIWG) is to coordinate government spending on File Systems and I/O (FSIO) R&D by all the government agencies that are involved in High End Computing. The HECIWG tasked a smaller advisory group to list, categorize, and prioritize HEC I/O and File Systems R&D needs. In 2005, leaders in FSIO from academia, industry and government agencies collaborated to list and prioritize areas of research in HEC FSIO. This led to a very successful High End Computing University Research Activity (HECURA) call from NSF in 2006 and has prompted a new HECURA call from NSF in 2009. This paper serves as both a review of the 2008 HEC FSIO identified research gaps as well as a preview of this forthcoming HECURA call.

Categories and Subject Descriptors
E.5 [**Files**]: Organization/structure; B.3.2 [**Memory Structures**]: Design Styles – *mass storage*; D.4.3 [**Operating Systems**]: File Systems Management – *access methods, directory structures, distributed file systems, file organization*; H.3.2 [**Information Storage and Retrieval**]: Information Storage – *file organization*.

General Terms
Algorithms, Management, Measurement, Performance, Design, Reliability, Security, Standardization

Keywords
File Systems, Storage, High End Computing

1. INTRODUCTION
The need for immense and rapidly increasing scale in high-end scientific computation drives the need for rapidly increasing scale in storage capability for scientific processing. Individual storage devices are rapidly getting denser while their bandwidth is not growing at the same pace. In the past several years, initial research into highly scalable file systems, high level Input/Output (I/O) libraries, and scalable I/O middleware was conducted to provide some solutions to the problems that arise from massively parallel storage. To help plan for the research needs in the area of File Systems and I/O, the inter-government-agency published "HPC File Systems and Scalable I/O: Suggested Research and Development Topics for the Fiscal 2005-2009 Time Frame"[2]. In turn, the High End Computing Interagency Working Group (HECIWG) designated this as an area of national focus starting in FY06. To collect a broader set of research needs in this area, the first HEC File Systems and I/O (FSIO) workshop was held in August 2005 in Grapevine, TX. Government agencies, top universities in the I/O area, and commercial entities that fund file systems and I/O research were invited to help the HECIWG determine the most needed research topics within this area.

The workshop attendees helped

- catalog existing government funded and other relevant research in this area,
- list top research areas that need to be addressed in the coming years,
- determine where gaps and overlaps exist, and
- recommend the most pressing future short and long term research areas and needs necessary to help advice the HEC to ensure a well coordinated set of government funded research

The recommended research topics are organized around these themes: metadata, measurement and understanding, quality of service, security, next-generation I/O architectures, communication and protocols, archive, and management and RAS. Additionally, university I/O center support in the forms of computing and simulation equipment availability, and availability of operational data to enable research, and HECIWG Institutions

involvement in the educational process were called out as areas needing assistance.

As a result of the information gathered at the 2005 workshop, the National Science Foundation issued a call to fund research into I/O, file and storage systems in the high-end computing environment under the HECURA program.

A second HECURA call focused on FSIO is scheduled for 2009 and will continue to fund R&D in this vital HEC technical area, because without a pipeline of R&D, breakthrough concepts will not emerge. The solicitation will concentrate on areas that are still considered to be gaps and will seek to round out the overall portfolio of R&D to cover the gaps as well as seek proposals which are both evolutionary and revolutionary.

2. Formation of the HEC FSIO Advisory Group

Shortly after the Japanese Earth Simulator Supercomputer came online and the US government realized that the US no longer had clear leadership in the supercomputing arena, the Presidents Information Technology Advisory Council formed the HEC Revitalization Task Force (RTF) which produced a set of recommendations[1]. One outcome of these recommendations was the formation of the HECIWG to find ways to coordinate government R&D to make a much more coordinated investment strategy for HEC. This HECIWG decided to pilot HEC Technical Advisory Groups (TAG) in a few important R&D areas to see if coordinating government funding of HEC R&D could help the nation become more competitive in high end computing. HEC FSIO was one of the areas they decided to pilot and so the HEC FSIO TAG was formed.

The HECIWG is composed of federal agencies including the Department of Commerce, the Department of Defense, the Department of Energy, the Department of Health and Human Services, the Department of Homeland Security, and the National Science Foundation. These agencies have funded tens of millions of dollars in dozens of HEC FSIO research projects at universities and national labs through the HECURA and other programs.

3. HEC FSIO Research Gap Areas

The areas in need of research identified at the HEC FSIO 2005 workshop were: metadata, measurement and understanding, quality of service (QoS), security, next-generation I/O architectures, communication and protocols, archive, and management and RAS. In the following sections, the gaps are examined and R&D roadmaps are provided. Each gap area is broken down into sub areas and each of these sub areas are ranked in three categories; how important is this to HEC R&D funding government agencies, the degree that research needs to be done in the area, and if there are products or ideas from the current R&D portfolio that are ready for commercialization.

3.1 Metadata

Metadata is usually described as data about data. In a file system context, metadata can be thought of as the information about a file and/or the data in a file that is stored external to the file. Permissions, atime, and mtime are all examples of metdata. Investigation into metadata issues is needed, especially in the areas of scalability, extensibility, access control, reliability, availability, and longevity for both file and archival systems. Scaling of metadata operations in both file systems and archives has not been fundamentally solved. While some engineering solutions have been put forth for scaling some metadata operations there are still no solutions that address this area fundamentally. Given the number of processing elements of future supercomputers, this is one of the top problems.

Additionally, consideration for very revolutionary ideas such as new approaches to name spaces and use of novel storage devices needs to be explored. Extensible metadata and alternative to tree based organization and access for files needs to be explored to enable indexed searches on file space on other than path name, for both users and administrators. Walking the file system tree even with parallel metadata ops may eventually not be feasible. There has been basically no progress in this alternative tree based metadata area.

Table 1. Metadata Road Map

Sub Areas	Rankings
Scaling	●▨▲ All existing work is evolutionary. What is lacking is revolutionary research; no fundamental solutions proposed. This category includes archive metadata scaling. More research in reliability at scale is needed.
Extensibility and Name Space	●■▲ All existing work is evolutionary. Extensibility includes provenance capture.
File System/ Archive Metadata Integration	○□△ Extended attributes, although not standardized, could solve problem.
Hybrid Devices Exploitation	●▨▲ Research is being done, but little is focused on metadata
Data Transparency and Access Methods	⊘▨△ No research focused on metadata
Importance: ● Very ⊘ Medium ○ Low Needs Research: ■ Greatly ▨ Medium □ Does Not Commercialization: ▲Greatly Needs ◭ Need and Ready for △ Not Ready for	

3.2 Measurement and Understanding

Measuring and understanding performance as more and more storage devices and clients are involved becomes difficult. Research into measurement and understanding of end-to-end I/O performance is needed including evolutionary ideas such as layered performance measurement, benchmarking, tracing, and visualization of I/O related performance data. Also, more radical ideas like end-to-end modeling and simulation of I/O stacks and

the use of virtual machines for large scale I/O simulation need to be explored.

Table 2. Measurement and Understanding Road Map

Sub Areas	Rankings
Understand System Workload in HEC Environments	● ▨ △ A comprehensive tool is nowhere in sight; problem is complex.
Standards and common practices for HEC I/O benchmarks and trace formats	⊘ □ △ Danger of over simplifying problem and could drive vendors to incorrect solutions.
Testbeds for I/O Research	● ▨ △ Simulators are being developed. No real testbeds being built. This problem will only get worse over time, i.e. as systems get bigger.
Applying cutting edge analysis tools to large scale I/O	● ■ △ Data are becoming available from Labs including I/O traces. Many opportunities to evaluate this research.
Importance: ● Very ⊘ Medium ○ Low Needs Research: ■ Greatly ▨ Medium □ Does Not Commercialization: ▲ Greatly Needs ◬ Need and Ready for △ Not Ready for	

3.3 Quality of Service

Quality of service (QoS) can be defined as features of a storage architecture that allow a user or administrator to recommend policies for data movement during I/O operations. These QoS policies can reach a broad range of integration into software, file systems, and hardware devices. Policies such as guaranteed I/O performance, specific redundancy requirements, or I/O priority settings will allow the system to perform optimally for a given work profile. Further research into areas such as adaptive QoS systems, end-to-end solutions, hardware support, and cross-system integration will revolutionize storage systems that will be created in the next few years is needed. These research topics will bring the storage systems to a point where users, systems, or entire clusters can be insulated from each other, while using the same storage infrastructure. This will also allow for predictable I/O performance and response time for the users.

The QoS work can not be considered complete until it addresses the question of how the user will interact with the QoS system, i.e. a standard API. The API must be easy for the user to express their needs and this is predicated on the assumption that the user will be able to easily identify their resource needs. Workloads with changing resource needs also provide additional challenges.

Table 3. Quality of Service Road Map

Sub Areas	Rankings
End to End QoS in HEC	⊘ ▨ ◬ Good research, but much work needed to get a standards based solution. Scale and dynamic environments have to be addressed at some point in time.
Standard Interfaces for QoS	⊘ ▨ △ Very partially addressed by proposed HEC POSIX Extensions. Will be driven by above "End to End QoS in HEC".
Importance: ● Very ⊘ Medium ○ Low Needs Research: ■ Greatly ▨ Medium □ Does Not Commercialization: ▲ Greatly Needs ◬ Need and Ready for △ Not Ready for	

3.4 Communication and Protocols

Communications and protocols will need to be used in these immensely scaled up environments. Protocols can be used to help in the huge orchestration effort. In the area of file system related communications and protocols, evolutionary items such as exploitation of Remote Direct Memory Access (RDMA), Object Based Secure Disk (OBSD) extensions, Network File System Version 4 (NFSv4) extensions, and parallel Network File System (pNFS) proof-of-concept implementations as well as more revolutionary exploration of server to server communications are needed.

Table 4. Communication and Protocols Road Map

Sub Areas	Rankings
Active Networks	⊘ ▨ △ Novel work being done, but not general enough.
Alternate I/O transport Schemes	⊘ □ △ Most aspects are being addressed.
Coherent Schemes	⊘ ▨ ◬ No consensus on how to do this correctly, but some solutions are in products
Importance: ● Very ⊘ Medium ○ Low Needs Research: ■ Greatly ▨ Medium □ Does Not Commercialization: ▲ Greatly Needs ◬ Need and Ready for △ Not Ready for	

3.5 Security

Security is an age-old issue, and while scale is one area that makes security difficult, there are many others as well especially need to know issues. Security for persistent data is difficult even without the scale of HEC systems. Aspects of security such as usability, long term key management, distributed authentication,

and dealing with security overhead are all topics for research. There is also room for more difficult research topics such as novel approaches to file system security including novel techniques for end-to-end encryption that can be managed easily over time. The need for standardization of access control list mechanisms is also needed and investigation into a standard API for end-to-end encryption could be very useful.

Table 5. Security Road Map

Sub Areas	Rankings
End-to-end encryption	○□△ Data is needed to validate designs
Performance overhead and distributed scaling	⊘▨◭ Problem is understood reasonably well. It is unclear if there is enough demand for commercialization.
Tracking of information	●▨△ Industry will help some, but not in HEC context. Includes tracking of provenance and flow information.
Ease of use, ease of management, quick recovery, ease of use API's	○□△ Data is needed to validate designs. Nothing to commercialize yet. Note: NSF should incorporate this into a call for security research; this topic is larger than FSIO.
Importance: ● Very ⊘ Medium ○ Low Needs Research: ■ Greatly ▨ Medium □ Does Not Commercialization: ▲Greatly Needs ◭ Need and Ready for △ Not Ready for	

3.6 Archive

The interest in archives is not in extending functionality or capabilities of archives, but focuses on integration of data flow from archives to file systems and I/O stacks. In the area of archive, the interfaces to the file systems and I/O stacks in HEC systems and long term care for the massive scale of an archive in the HEC environment are difficult areas needing more research than they have received previously.

Table 6. Archive Road Map

Sub Areas	Rankings
API's/Standards for interface, searches, and attributes, staging etc.	⊘▨△ Current research is in terms of file systems, not archive. API merging with POSIX and API for searching lacking
Long term attribute driven security	○▨△ Current research is in terms of file systems, not archive. Current researchers need data supporting proposed solutions usefulness
Long term data reliability and management	●□◭ Need for research and commercialization is low because HIPPA and others will drive research/products. Redundancy techniques reasonably sufficient for archives
Policy driven management	○□◭ Sarbanes-Oxley Act is solving this problem
Importance: ● Very ⊘ Medium ○ Low Needs Research: ■ Greatly ▨ Medium □ Does Not Commercialization: ▲Greatly Needs ◭ Need and Ready for △ Not Ready for	

3.7 Management and RAS

Management and RAS in an environment with 100,000 or more disks and million way parallelism will be extremely difficult. Adding to the difficulty of management is the added difficulty of having to deal with long term persistent data. Failure of processors typically means a job is re-run. Failure of a storage device may mean loss of valuable data or information. In the area of management, reliability and availability at scale, management scaling, continuous versioning, and power management of storage systems are all needed research topics. Additionally, more revolutionary ideas like autonomics, use of virtual machines, and novel devices exploitation need to be explored.

Table 7. Management and RAS

Sub Areas	Rankings
Automated problem analysis and modeling	⊘▨△ More researchers need to look at this problem
Formal Failure analysis and tools for storage systems	⊘□◭ Good research done here. Will people use this work?
Improved Scalability	⊘▨△ More research is needed here. Test beds are probably needed for this work.
Power Consumption and Efficiency	○▨◭ Industry is working on this problem. Storage is not a large consumer of energy at HEC sites.
Reliability, and degraded performance in HEC systems	⊘▨◭ Industry is working on this problem
Importance: ● Very ⊘ Medium ○ Low Needs Research: ■ Greatly ▨ Medium □ Does Not Commercialization: ▲Greatly Needs ◭ Need and Ready for △ Not Ready for	

3.8 Next-Generation I/O Architectures

Next generation I/O architectures will be needed to orchestrate movement of data and metadata from so many processing elements and so many disks in our HEC environments. There is great need for research into next-generation I/O architectures, including evolutionary concepts such as extending the POSIX I/O API standard to support archives in a more natural way, access awareness, and high concurrence at HEC scale. Studies into methods to deal with small, unaligned I/O and mixed-size I/O workloads as well as collaborative caching and impedance matching are also needed. Novel approaches to I/O and File Systems also need to be explored including redistribution of intelligence, adaptive and reconfigurable I/O stacks, user space file systems, data-aware file systems, and the use of novel storage devices.

Table 8. Next Generation I/O Architectures Road Map

Sub Areas	Rankings
Understanding file system abstractions - Scalable file system architectures	●■◬ Good work, but much of the research is in its infancy. A small portion ready for commercialization.
Self-assembling, Self-reconfiguration, Self-healing storage components	●■△ Good work being done, but it's a hard problem that will take more time to solve.
Hybrid architectures leveraging emerging storage technologies	●▧◬ Big potential reward, but very little work being done in the HEC area. Includes power consumption. Traditional block-based solutions ready for commercialization. Alternative interfaces not yet well explored.
HEC systems with multi-million way parallelism doing small I/O Operations	⊘▧◬ Good initial research; needs to be moved into testing. More fundamental solutions being pondered including non-volatile solid state storage
Importance: ● Very ⊘ Medium ○ Low Needs Research: ■ Greatly ▧ Medium □ Does Not Commercialization: ▲Greatly Needs ◬ Need and Ready for △ Not Ready for	

3.9 Assisting with Standards, Research, and Education

At the HEC FSIO 2005 workshop, there was a recognition that the HEC FSIO community should find ways of supporting students working in the general area of I/O as well as students working more specifically on I/O within HEC. Investment to support the research of these students was considered worthwhile both because they may provide important research while still in school as well as by cultivating these students such that they may continue to work on HEC I/O problems following their graduation and, with any luck, become the next generation of HEC I/O experts.

Over the past decade, the HEC community has had a role in the formation and adoption of various FSIO related standards. The most notable are the ANSI T10 1355D specification for Object Based Storage Devices (OBSD), the IETF NFSv4 standard including the new pNFS portion of the NFSv4.1 minor revision of the NFSv4 specification. The newly formed Open Group HEC Extensions to the POSIX standards work has also been an outcome of HEC FSIO and the HEC I/O community work

Over the past few years many HEC sites have released data in failure, operational, and usage of supercomputers, and I/O traces or synthetic and real application workloads. The data and tools have been a huge help to the community, but more are needed. We also need to ensure distribution and acceptance and usage of the tools. These tools and this data can assist in future designs in reliability and performance and used to validate researchers models.

4. Conclusion

Scalable I/O is perhaps the most overlooked area of HEC R&D. Given the information generating and processing capabilities being installed and contemplated at HEC sites, it is a mistake to continue to neglect this area of HEC. One of the primary purposes of this document is to present the areas in need of new and continued investment in R&D and standardization in this crucial area of HPC file systems and scalable I/O that should be pursued by the government.

In the near future, sites will routinely deploy supercomputers with hundreds of thousands processors. Million-way parallelism is around the corner and, with it, storage bandwidth requirements will go from tens of gigabytes per second to terabytes per second. Online storage requirements to support work flows for efficient complex science will begin to approach the exabyte range. The ability to handle a more varied I/O workload ranging seven orders of magnitude in performance characteristics, extremely high metadata activities, and management of trillions of files will be required. Global or virtual enterprise wide-area sharing of data with flexible and effective security will be required. Current extreme-scale file system deployments already suffer from reliability and availability issues, including recovery times from corruption issues and rebuild times. As these extreme-scale deployments grow larger, these issues will only get worse. It will possibly be unthinkable for a site to run a file system check utility, yet it is almost a given that corruption issues will arise. Recovery times need to be reduced by orders of magnitude, and these types of tools need to be reliable, even though they may rarely be used. The number of storage devices needed in a single coordinated operation could be in the tens to hundreds of thousands, requiring integrity and reliability schemes that are far more scalable than available today. Management of enterprise-class global parallel file/storage systems will become increasingly difficult due to the number of elements involved, which will

likely approach 100,000 spinning disks with widely varying workloads. The challenges of the future are formidable.

5. ACKNOWLEDGMENTS

The HECFSIO Technical Advisory Committee would like to thank Almadena Y. Chtchelkanova, Program Director of Computing and Communication Foundations (CISE/CCF) at the National Science Foundation for all of her efforts in promoting the need for R&D funding in High End Computing and HEC FSIO.

Thanks to the all of the organizations that compose the HECIWG and their recent leaders including Bob Meisner and Thuc Hoang of DOE/NNSA ASC, Fred Johnson of DOE/Office of Science, and John Grosh formerly of DARPA.

In addition to all the active members of the Technical Advisory Group who donate their time and efforts, we'd also like to thank a past member of the advisory team, Bill Loewe, formerly of Lawrence Livermore National Lab, for his contribution to this work and the HEC FSIO effort.

This work is partially funded by the SciDAC Petascale Data Storage Institute; DOE award DEFC0206ER25767.

6. REFERENCES

[1] NITRD High End Computing Revitalization Task Force (HECRTF). 2003. *Report of the Workshop on the Roadmap for the Revitalization of High-End Computing*. Daniel A. Reed, ed. June 16-20, Washington, D.C.

[2] Ross, R., Felix, E., Loewe, B., Ward, L., Grider, G., and Hill, R. 2005 *HPC File Systems and Scalable I/O: Suggested Research and Development Topics for the Fiscal 2005-2009 Time Frame*, tech. report, 2005, http://institutes.lanl.gov/hec-fsio/docs/FileSystems-DTS-SIO-FY05-FY09-R&D-topics-final.pdf